ECE 6882

HW #2

Date:

Osama Yousuf

P1

$R(a^1) \sim \text{Uniform}[0 \quad 1.4]$
$R(a^2) \sim N(\mu = 0.5, \sigma = 1)$

a) Compute $Q^*(a^1), Q^*(a^2), \pi^*$.

$Q^*(a^1) = E[R(a^1)] = \frac{0+1.4}{2}$

$\Rightarrow \boxed{Q^*(a^1) = 0.7}$

$Q^*(a^2) = E[R(a^2)] = \mu = 0.5$

$\Rightarrow \boxed{Q^*(a^2) = 0.5}$

Since $Q^*(a^1) > Q^*(a^2)$, optimal policy is simply,

$\pi^* = \text{argmax}\, Q(a^i) = a^1$

$\Rightarrow \boxed{\pi^* = a^1}$

Date: ____________

b) $\alpha = 0.5$

I have assumed that initial values $Q(a^1) = Q(a^2) = 0$ for this part.

When action a is selected at time k, corresponding Q-value is updated by:

$$Q(a) := Q(a) + \alpha\left(r - Q(a)\right)$$

$k=1$

a^1 is chosen, $r=1$

$$Q(a^1) = 0 + 0.5(1-0) = 0.5$$

$$\Rightarrow \boxed{Q(a^1) = 0.5,\ Q(a^2) = 0}$$

$k=2$:

a^2 is chosen, $r = 0.5$

$$Q(a^2) = 0 + 0.5(0.5 - 0) = 0.25$$

$$\Rightarrow \boxed{Q(a^1) = 0.5,\ Q(a^2) = 0.25}$$

$k=3$:

a^1 is chosen, $r = 0$

$$Q(a^1) = 0.5 + 0.5(0 - 0.5) = 0.25$$

$$\Rightarrow \boxed{Q(a^1) = Q(a^2) = 0.25}$$

Date: ______

$k=4$

a^2 is chosen, $r=1.25$

$Q(a^2) = 0.25 + 0.5(1.25-0.25)$

$\boxed{Q(a^2) = 0.75}$ $\boxed{Q(a_1) = 0.25}$

$k=5$:

a^1 is chosen, $r = 1.35$

$Q(a^1) = 0.25 + 0.5(1.35 - 0.25)$

$\boxed{Q(a^1) = 0.8, \; Q(a^2) = 0.75}$

Since $Q(a^1) > Q(a^2)$,

$\boxed{\pi = a^1}$

c) We repeat b, assuming $Q(a^1) = Q(a^2) = 5$.

$k=1$:

$Q(a^1) = 5 + 0.5(1 - 5) = 3$

$\Rightarrow Q(a^1) = 3, \; Q(a^2) = 5$

$k=2$:

$Q(a^2) = 5 + 0.5(0.5 - 5) = 2.75$

$\Rightarrow Q(a^1) = 3, \; Q(a^2) = 2.75$

Date: ____________

$k=3$:

$$Q(a^1) = 3 + 0.5(0-3) = 1.5$$

$$Q(a^1) = 1.5, \quad Q(a^2) = 2.75$$

$k=4$:

$$Q(a^2) = 2.75 + 0.5(1.25 - 2.75) = 2$$

$$Q(a^1) = 1.5, \quad Q(a^2) = 2$$

$k=5$:

$$Q(a^1) = 1.5 + 0.5(1.35 - 1.5) = 1.425$$

$$\boxed{Q(a^1) = 1.425, \quad Q(a^2) = 2}$$

$$\boxed{\pi = a^2} \quad (\text{since } Q(a^2) > Q(a^1))$$

We conclude that the optimistic initial Q values perform a higher degree of exploration and are therefore unable to determine π^* under given parameters & time steps.

Date: ______

P2 $\alpha = 0.5$, $H_1(a^1) = H_1(a^2) = 0$.

→ $k=1$

a^1 chosen, $R_t = R_1 = 1$, $\bar{R}_t = 1$

$$\pi_1(a^1) = \frac{e^{H_1(a_1)}}{e^{H_1(a^1)} + e^{H_1(a^2)}} = \frac{1}{2}$$

$$\pi_1(a^2) = \frac{e^{H_1(a^2)}}{e^{H_1(a^1)} + e^{H_1(a^2)}} = \frac{1}{2}$$

$$\Rightarrow \boxed{\pi_1(a_1) = \pi_1(a_2) = 1/2 = 0.5}$$

New preferences, ($\bar{R}_t$ = avg. so far)

$$H_{t+1}(a_t) = H_t(a_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(a_t))$$

for selected action.

$$\Rightarrow H_2(a_1) = H_1(a_1) + 0.5(1-1)(1-\pi_1(a_1))$$

$$\Rightarrow \boxed{H_2(a_1) = 0}$$

$$H_{t+1}(a) = H_t(a) - \alpha(R_t - \bar{R}_t)\pi_t(a)$$

for unselected action

$$\Rightarrow H_2(a_2) = H_1(a_2) - 0.5(1-1)\pi_1(a_2)$$

$$\Rightarrow \boxed{H_2(a_2) = 0}$$

Date: ____________

→ $\underline{k=2}$:

a^2 chosen, $R_t = R_2 = 0.5$

Avg. reward so far $= \bar{R}_t = \frac{1+0.5}{2} = 0.75$

$$\boxed{\pi_2(a^1) = \pi_2(a^2) = 1/2 = 0.5}$$

(since $H_2(a_i) = H_1(a_i)$)

Chosen a_2:

$$H_{2+1}(a_2) = H_2(a_2) + \alpha(0.5 - 0.75) \times (1 - \pi_2(a_2))$$

$$H_3(a_2) = 0 + (0.5 \times -0.25 \times 0.5)$$

$$\boxed{H_3(a_2) = -0.0625}$$

Unchosen a_1:

$$H_3(a_1) = H_2(a_1) - \alpha(0.5 - 0.75)\pi_2(a_1)$$

$$H_3(a_1) = 0 - 0.5 \times -0.25 \times 0.5$$

$$\boxed{H_3(a_1) = +0.0625}$$

→ $\underline{k=3}$: a_1 chosen, $R_3 = 0$, $\bar{R}_t = \frac{1+0.5+0}{3}$

$\bar{R}_t = 0.5$

$$\pi_3(a_1) = e^{0.0625} / (e^{0.0625} + e^{-0.0625})$$

$$= 0.531$$

$$\pi_3(a_2) = e^{-0.0625} / (e^{0.0625} + e^{-0.0625})$$

$$= 0.469$$

Chosen a_1:

$$H_4(a_1) = H_3(a_1) + \alpha(0 - 0.5)(1 - \pi_3(a_1))$$

$$H_4(a_1) = 0.0625 + 0.5(-0.5)(1 - 0.531)$$

$$\boxed{H_4(a_1) = -0.05475}$$

Date:

Unchosen a_2:

$$H_4(a_2) = H_3(a_2) - \alpha(R_3 - \bar{R}_3)(\pi_3(a_2))$$
$$= -0.0625 - 0.5(0 - 0.5)(0.469)$$
$$\boxed{H_4(a_2) = +0.05475}$$

Finally for $\pi_4(a^i)$'s,

$$\pi_4(a_1) = e^{-0.05475} / (e^{0.05475} + e^{-0.05475})$$
$$\boxed{\pi_4(a_1) = 0.4726}$$
$$\pi_4(a_2) = e^{0.05475} / (e^{0.05475} + e^{-0.05475})$$
$$\boxed{\pi_4(a_2) = 0.5274}$$

Thus,

$$H_4(a_1) = -0.05475$$
$$H_4(a_2) = +0.05475$$
$$\pi_4(a_1) = 47.26\%$$
$$\pi_4(a_2) = 52.74\%$$

At timestep $k = 4$, a_2 will have a slightly higher preference.

x ——— x ——— x